University of Colorado
Boulder

Calculating RNA motif probabilities and recognizing patterns in sequence data

A thesis submitted in partial satisfaction
of the requirements for the degree
Bachelors of Science in Computer Science

by

Ryan Kennedy

2009

The thesis of Ryan Kennedy is approved.

Michael Mozer

Manuel Lladser, Committee Co-chair

Rob Knight, Committee Chair

University of Colorado, Boulder

2009

Table of Contents

1 Background **1**

1.1 Nucleic Acids 2

1.1.1 RNA Motifs 3

1.2 Transcription and transcription factors 5

2 Probabilistic Analysis of RNA **7**

2.1 Motivation 7

2.2 Previous Work 8

2.2.1 Exact Calculation 8

2.2.2 Information Content Approximation 12

2.2.3 Poisson Approximation 13

2.2.4 Upper Bound Approximation 14

2.3 Approximation with Confidence Bounds 16

2.3.1 Determining the normality of the distribution 18

2.3.2 Extension to Markovian backgrounds 19

3 Application to the Analysis of RNA Motifs **21**

4 Searching for Patterns in Sequence Data **26**

4.0.3 Background 26

4.0.4 Effect of the Sequence 27

4.0.5 Visualization of the data sets 28

4.0.6 Supervised classification 31
4.0.7 Detecting covariation in sequences 36
4.0.8 Effect of Context 41

5 Discussion 46

References 50

List of Figures

1.1 A hairpin secondary structure in an RNA molecule. The red region forms a loop who's bases maybe be arbitrary and of any number for the motif to be functional. 5

2.1 An example of a deterministic finite automaton that recognizes any string containing the substring 'ACCG' 9

2.2 Computation time for simple motifs as the number of degenerate correlations increases. Both the information content method (blue) and the Poisson approximation (red) scale very well with increased motif complexity. In contrast, the exact calculation (black) grows fast and the calculation is infeasible for more than three correlations. The asymptotic confidence bound method (green) scales much better than the exact method and remains computationally feasible for complex motifs. 18

3.1 Distribution of one-standard-deviation ellipsoids for Sequence, Folding and Sequence&Folding 22

3.2 Distribution of sequences taken from Rfam, by function. 23

3.3 Comparison of distributions of (a) natural aptamers (green) and ribozymes (blue) and (b) artificially selected motifs. The two distributions are superimposed in (c). 25

4.1 PCoA for several U.S. cities. 30

4.2 PCoA for positive control. 31

4.3 PCoA for Horvath data set using Hamming distance. 31

4.4 PCoA for Horvath data set using alternative distance. 32
4.5 PCoA for Riley data set using Levenshtein distance. 32
4.6 Secondary structure of the histone3 RNA. Adapted from Rfam. . 38
4.7 Mutual information heatmap for histone3. 39
4.8 Mutual infomation heatmap for Horvath - full alignment. 40
4.9 Mutual infomation heatmap for Horvath - apoptosis sites. 41
4.10 Mutual infomation heatmap for Horvath - cell cycle sites. 41
4.11 Mutual infomation heatmap for Riley - full alignment. 42
4.12 Mutual infomation heatmap for Riley - activators. 42
4.13 Mutual infomation heatmap for Riley - repressors. 43
4.14 Distributions of p53 log-odds matrix scores for verified binding sites and random sequences. 45

List of Tables

1.1 IUPAC specification of degenerate RNA bases. 4

4.1 Distance matrix for several U.S. cities, in miles. Data from [1]. . . 29

4.2 Confusion matrix for positive control using the Hamming distance for random forest classification. 33

4.3 Confusion matrix for Horvath data set (7 classes) using the Hamming distance for random forest classification. Class names are abbreviated for space. 34

4.4 Confusion matrix for Horvath data set (2 classes) using the Hamming distance for random forest classification. 34

4.5 Confusion matrix for Riley data set using the Levenshtein distance for random forest classification. 35

4.6 Confusion matrix for positive control using the Hamming distance for random forest classification, using the sequence data directly. . 36

4.7 Confusion matrix for Horvath data set (7 classes) using the Hamming distance for random forest classification, using the sequence data directly. Class names are abbreviated for space. 37

4.8 Confusion matrix for Horvath data set (2 classes) using the Hamming distance for random forest classification, using the sequence data directly. 37

4.9 Probability weight matrix for STAT1 44

Acknowledgments

I would first like to thank my advisors, Rob Knight and Manuel Lladser. Working with both of them for the past couple years has been an incredible experience and they have given me an enormous amount of guidance and support. Without their help, I would surely not be where I am today; it was because of them that I first became interested in research. Working with them has been a pleasure and I have been very lucky to have had the opportunity. I'd like to thank them also for taking time to provide feedback on this thesis. I'd also like to thank the Knight lab in general, who have always provided a warm and supportive atmosphere. In particular, I'd like to thank to Daniel McDonald and Micah Hamady, who have both helped me run many scripts on the department computer cluster. The other member of my thesis committee, Michael Mozer, deserves a lot of thanks as well. Since working with him on a research project my junior year, he has been supportive of all my work and his feedback has been very much appreciated.

In support of my research, I want to thank Michael Yarus. Also, thanks to Joaquin Espinosa, who helped with my analysis of p53 in Chapter 4. I would also like to acknowledge Hans De Sterck and Zhiyuan Wu at the University of Waterloo, who I have collaborated with, along with Rob Knight and Manuel Lladser, for the work presented in Chapter 3.

Another person who has had a tremendous influence on my work is Anne Dougherty, my academic advisor. She has guided me down the right path countless times and a better advisor would be hard to find. Jim Curry, chair of the Applied Math department, has also supported me in my research and for this I thank him as well.

I'd also like to thank both of my parents, Gene and Sarah Kennedy, for their

support of me. From making time to attend my presentations to attempting to decipher my research papers, they have always been source of encouragement. Of course, it's too little to thank them for just this, since this encouragement has been present for the last 22 years. Thanks also to my brother, Shane, for his support. I would also like to thank my girlfriend, JJ, who has been a constant source of encouragement and strength.

Much of this thesis has also been aided by generous scholarships, including an Astronaut Foundation Scholarship, Norlin Scholarship, Robert C. Byrd Memorial Scholarship, Frank J. LaRocca Memorial Scholarship, and an Applied Math Undergraduate Scholarship. The support of these awards has allowed me to focus on the research presented in this thesis and is very much appreciated.

Publications

Information, probability, and the abundance of the simplest RNA active sites. Ryan Kennedy, Manuel E. Lladser, Michael Yarus, Rob Knight. Frontiers in Bioscience 13, 6060-6071, May 1, 2008

Active site specifications and self-organization drive universal compositional biases in naturally and artificially selected RNA sequences. Ryan Kennedy, Manuel Lladser, Zhiyuan Wu, Chen Zhang, Michael Yarus, Hans De Sterck, Rob Knight. Manuscript in preparation for submission to *Nature*.

Author's Contributions

The techniques presented in Chapters 2 and 3 of this thesis were conceived of by Rob Knight and Manuel Lladser. Under their supervision, I implemented a library written in Python which allows for the calculation of all of the techniques described therein. For our first publication which used these techniques, [18], I performed the analyses using this software and created all of the figures. For the application of these techniques, presented in Chapter 4, I also performed all of the analyses other than calculating the motif folding probabilities. These analyses were extensive and consisted of writing code to organize, extract, and display the data in different ways. Also, I performed all of the statistical testing required for our second paper [17]. The work presented in Chapter 4 was guided by Rob Knight, although I performed all of the data collection and analyses.

Abstract of the Thesis

Calculating RNA motif probabilities and recognizing patterns in sequence data

by

Ryan Kennedy

Bachelors of Science in Computer Science

University of Colorado, Boulder, 2009

Professor Rob Knight, Chair

Recent advances in technology have catalyzed an explosion in the amount of available biological sequence data. In order to efficiently extract information from this vast amount of data, analyses using computers are essential. In this thesis, I present several techniques for the analysis of nucleic acids which allow for efficient research into many biological processes.

First, I look at the probability of a specific RNA motif occurring in a random sequence. While previous work on this topic has resulted in approximations to this problem with an unquantifiable amount error, I present a novel method that provides a confidence interval around the true probability which scales much better than exact calculations. I also compare the accuracy of different methods and find that some approximations are accurate across a wide range of conditions while others are frequently in error by several orders of magnitude. I apply these techniques to an analysis of biological RNA motifs in order to explore compositional biases found in both naturally-occurring and artificially-selected

motifs.

In addition, I present a framework of machine learning techniques for extracting information from sequence data in molecular biology. This framework is able to identify interactions relating to differences within individual sequences - including covariation between positions - as well as interactions between the sequences themselves. I demonstrate the use of these methods in looking for patterns related to the transcription factor p53, a protein known to have an effect on tumor suppression.

CHAPTER 1

Background

Over the last few years, the amount of publicly available biological data has undergone an enormous increase. For example, GenBank [2] - a government-maintained database of publicly available DNA sequences - had only 606 sequences in 1982. The number of sequences reached a million in 1996 and has continued to grow exponentially; as of 2008, GenBank contains nearly 100 million DNA sequences, or about 100 billion base pairs. When the amount of available data was much less, extracting information from it was possible by visual inspection or manual manipulation. Now, however, advances in technology allow for millions of sequences to be produced very quickly, necessitating the use of computers. Computers also provide the additional advantage of being able to model data and produce results much more quickly and efficiently than ever before. Even so, this is not done automatically. New analysis techniques must be designed in order to extract this information from data. In this thesis, I explore several different analyses of nucleic acids and how they can be applied to different data sets.

The basis for the work presented in this thesis lies in biology, and so this section will provide a basic introduction to the biological concepts used in the

rest of the thesis. Much of the thesis is quantitative in nature, however, and can be understood without having a firm grasp of biology. Even so, some of the conclusions that are drawn, as well as the motivations behind the work, come from biology and so here is provided a brief background.

1.1 Nucleic Acids

Nucleic acids are a family of molecules to which DNA and RNA belong [12]. DNA - deoxyribonucleic acid - is commonly known as the carrier of a person's genetic information and is present in every cell in the body. DNA is composed of four types of nucleotide bases: adenine, cytosine, guanine and thymine, although they will hereafter be referred to as just A, C, G and T, respectively.

RNA - ribonucleic acid - is related to DNA and is also a carrier of genetic information. It too is composed of four bases, although DNA's thymine base is replaced by uracil (U) in RNA. There are several different types of RNA and together they play a wide range of roles. For example, RNA can be an intermediate carrier of genetic information or even have catalytic functions (these RNA enzymes are known as *ribozymes*) similar to those of proteins [12]. Because of these different functions that RNA can perform, it has been hypothesized that, before the existence of DNA or proteins, RNA existed in their place. This idea is known as the *RNA World* hypothesis [14].

Even though nucleic acids are composed of only four bases, they are much more complex than their four-base composition would suggest. Although they

can be described in terms of their *primary structure* - the sequences of bases from which they are made - these bases are able to pair together into more complex arrangements, known as their *secondary structure*. Standard Watson-Crick base pairing [12] tells us that the bases A and U (or T, for DNA) will pair together, as will the bases C and G. There is also the possibility of G-U base pair in RNAs, known as *wobble* base pairing [35], although our results have shown that this has little if any effect on the applications of the techniques that I explore in this thesis. Finally, nucleic acids are able to go beyond simple base pairing and fold into more complex shapes, known as the *tertiary structure*. As I have shown in [17] and will explore in this thesis, the composition of RNAs has a significant impact on whether they are able to fold into specific tertiary structures and this distribution gives functional RNAs certain compositional biases.

1.1.1 RNA Motifs

RNA *motifs* are specific patterns of RNA that are of interest. These motifs correspond to the minimal specifications of molecules that may perform a specific function. For example, the isoleucine aptamer [21] is an RNA that binds to the amino acid isoleucine. Aptamers are important in fields such as drug research, where aptamers which bind to deleterious molecules might be selected for. Other RNA motifs include self-cleaving ribozymes [20] and aptamers which bind to antibiotics [5].

The specifications for these patterns include the primary structure of the motif using the four bases A, C, G and U, as well as *degenerate* bases which may

represent several of the bases. For example, the degeneracy N corresponds to a base that is allowed to be any of the four bases while a Y represents the bases C and U. Table 1.1 lists the full IUPAC representation of degenerate bases in RNA [31].

Symbol	Bases Represented
A	A
C	C
G	G
U	U
W	A/U
S	C/G
M	A/C
K	G/U
R	A/G
Y	C/U
B	C/G/U
D	A/G/U
H	A/C/U
V	A/C/G
N	A/C/G/U

Table 1.1: IUPAC specification of degenerate RNA bases.

Motif specifications can also contain base pairings as well as spaces that separate the motif into *modules*. Modules are pieces of the motif that are specified such that any number of arbitrary bases can be put between the two module without changing the motif. This is frequently seen in secondary structures such

as the *hairpin*, (Figure 1.1) where two sections of RNA are specified while what is in between the two sections is of no consequence to the RNA's function. Much of the work in this thesis relates to RNA motifs, where we are looking at properties of specific, biologically-significant RNAs.

Figure 1.1: A hairpin secondary structure in an RNA molecule. The red region forms a loop who's bases maybe be arbitrary and of any number for the motif to be functional.

1.2 Transcription and transcription factors

DNA, the carrier of genetic information, resides in the nucleus of eukaryotic cells [12]. RNA is created by copying specific portions of the DNA molecule. This process of creating RNA from DNA is known as *transcription*. This process is important because transcribed messenger-RNA (mRNA) will go on to create proteins which play a central role in virtually every process in our body.

Transcription factors are proteins that play an important role in regulating transcription. These transcription factors bind to the DNA molecule and have some effect on transcription, possibly increasing or decreasing the rate of transcription. However, the interactions between transcription factors and the

regulation of transcription is complex, and for this reason I also explore a framework for extracting information about these interactions as applied to p53, a transcription factor that is known to have an effect on tumor suppression [16,32].

CHAPTER 2

Probabilistic Analysis of RNA

2.1 Motivation

Determining a hierarchy of how species have evolved is a problem which must be inferred *a posteriori*; it is not possible to go back in time to see the process of evolution itself. As a result, inferences have to be made from existing data, possibly including body structures of the animals or their genetic information. This same problem can be thought of for RNA, where specific RNA structures are found in a diverse group of animals (for example, the Hammerhead motif [4]). Unfortunately, it is difficult to distinguish between RNA motifs that evolved independently multiple times or evolved once and were passed on in subsequent generations. One way to make an inference is to calculate how likely it is for a motif to occur in a random sequence, giving an indication of how likely it is that the motif would have evolved independently. If a motif is very likely to evolve by itself, then it is probable that most of the observations of that motif arose independently. On the other hand, for highly-unlikely motifs, it is more probable that it evolved only a few times and was passed on to others. In this

chapter, I compare several different methods for calculating the probability of a motif occurring in a random sequence and also present a novel method with several advantages over previous work.

2.2 Previous Work

2.2.1 Exact Calculation

The most obvious method for calculating the probability of the occurrence of a given RNA motif is to calculate it exactly. The exact calculation can be accomplished by using certain types of deterministic finite automata (DFA) (Figure 2.1) [34]. A DFA is a set of states where, given a string, each character will cause a transition to another state of the automata. A string is said to be *accepted* by the DFA if the final state where it ends up is one of the accepting states of the automaton. An important characteristic of DFAs is that the set of languages that can be represented by a DFA is equivalent to the set of regular languages. This fact will an be important as we calculate the probability of motif occurrence using DFAs.

Because our RNA motifs are just strings, we can treat them as words and use a special class of DFAs to detect them: Aho-Corasick automata [10]. Aho-Corasick automata are used frequently in string matching and provide the ability to detect all matches of a given word in a string of characters; in our case, we are looking for an RNA motif in a longer sequence of nucleotide bases. However,

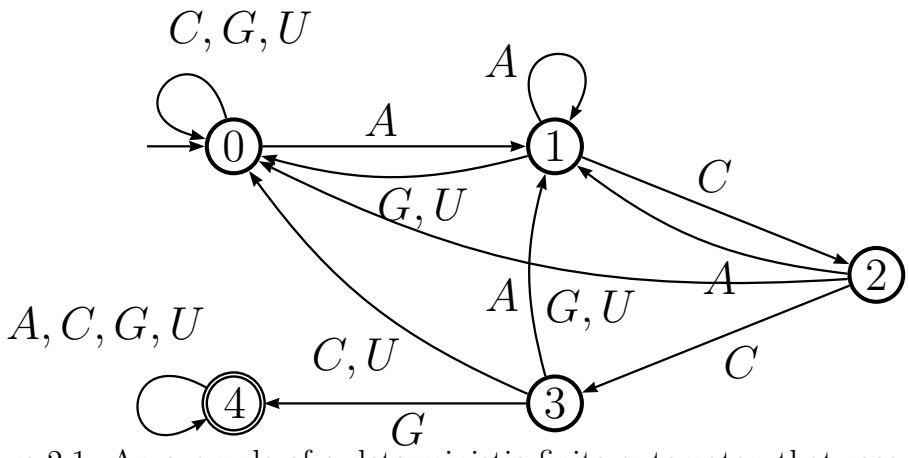

Figure 2.1: An example of a deterministic finite automaton that recognizes any string containing the substring 'ACCG'

because we want to detect whether or not the motif is present in the string at all and not just whether we saw the motif at the last base, we make the terminal states of the automata *absorbent*, meaning that any additional characters fed to the automata will not result in a transition to any other states (see Figure 2.1 for an example).

Although we can now detect the presence of an RNA motif by using an Aho-Corasick automaton, our eventual goal is determine the *probability* of a given motif occurring in a random string. Even though it would be possible to estimate this via Monte Carlo simulation, whereby many random strings are generated and the probability is estimated as the proportion of strings that contain the motif, this is impractical since we typically find that the true probability of seeing the motif is less than - sometimes much less than - 10^{-10}. Instead, we assign a probability to each transition in the automaton based on the probability of each nucleotide base in the random string. The resulting automaton can be represented as a first-order homogeneous Markov chain [24] where the transition matrix is now the probability transition matrix. Now, the probability of the motif represented

by the automaton occurring in a random string of length n is the probability of starting at the initial state and ending at a terminal state. This can be calculated by taking $\sum_{t \epsilon T} P^n(i,t)$, where i is the initial state, T is the set of terminal states and P is the probability transition matrix.

Using this method, it is indeed feasible to calculate the probability of occurrence of simple motifs which can be represented as a regular language. Consider, for instance, a simple motif with only a single module and no degenerate bases, such as `AGGUAUCAGUA`. Even motifs containing paired, non-degenerate bases or non-paired, degenerate bases can be represented in this form since pairing of non-degenerate bases does not change the motif at all and non-paired degenerate bases can be represented as a single state where transition probabilities are calculated by taking into account all possible bases the degeneracy represents. Furthermore, this can also represent motifs containing more than one module (where the number of bases between modules is $> k$) by taking the concatenation of automata that represent each module (with additional *spacer* states for $k > 0$). For these cases, calculation of the exact probability is typically not difficult.

Problems arise, however, because many real-world motifs contain paired degenerate bases (For example, the Tryptophan aptamer motif can be represented as `(N(N(R(Y(YRAGUNU(C(GCAGUAACC)G)URNN)G)Y)N)N)` [25], where matching parentheses surround paired bases). In order to apply this analysis of RNA to real motifs without imposing severe restrictions on the motifs themselves, it is important to be able to calculate the probability for this case. Unfortunately, the co-dependence between bases that is introduced through pairing of degenerate bases results in a language that is no longer regular and cannot be represented

by a single, simple finite automaton. Instead, we can represent such motifs by the entire set of possible non-degenerate motifs which they represent. For example, the motif `ACG(NAN)U` can be decomposed into the set of motifs {`ACGAAUU`, `ACGCAGU`, `ACGGACU`, `ACGUAAU`}. The exact probability of occurrence can then be calculated by finding the probability of occurrence for each of the individual motifs as previously described and then using the inclusion-exclusion principle, which says that

$$\text{Prob(motif)} = S_1 + S_2 + \ldots + S_m, \tag{2.1}$$

where, if M_i is the event that the i^{th} motif occurs, then

$$\begin{aligned} S_1 &= \sum_j \text{Prob}\left(M_j\right) \\ S_2 &= -\sum_{i<j} \text{Prob}\left(M_i \cap M_j\right) \\ &\cdots \\ S_k &= (-1)^{(k+1)} \cdot \sum_{I \subset \{1,\ldots,m\}:|I|=k} \text{Prob}\left(\cap_{i \epsilon I} M_i\right) \end{aligned}$$

Here, $\text{Prob}(\cap_i M_i)$ can be calculated using the product automaton of each individual motif [24]. The problem with this technique is that the complexity of the calculation grows very quickly with the number of degenerate base pairs. Since every set of fully-degenerate paired bases can be represented by four non-degenerate motifs, a motif with k fully-degenerate base pairs requires the inclusion-exclusion principle to be applied to a set of 4^k motifs. So, a relatively-small motif with just

4 such base pairs will result in a set of representative motifs with $4^4 = 256$ elements. In calculating the probability, then, the inclusion-exclusion rule's largest sum will contain $\binom{256}{256/2} \approx 5.8 \times 10^{75}$ terms. Clearly, this is far too many terms to calculate. Furthermore, real motifs are frequently much more complex than this; the Chloramphenicol motif [19] has 21 fully-degenerate base pairs.

2.2.2 Information Content Approximation

A relatively simple approximation to the true probability of motif occurrence can be calculated using the information content of a motif. This technique is used when a multiple sequence alignment of the RNA sequences is constructed and conserved positions (positions with the same nucleotide base in different sequences) are identified. Then, by calculating the Shannon entropy [36] ($I = -\Delta H, \text{where } H = -\sum p_i \ log_2 \ p_i$) of each position/base pair (both independent positions and base pairs have the same information content since each has four possible states, assuming no wobble pairing) and summing across the entire sequence, the resulting value is an indication of the amount of information that the motif contains. The reduction in entropy of each base/base pair - where all four possible states are equally-likely - is then

$$\Delta H = -\left(0 - \left(-4 \times 0.25 \times log_2 \ 0.25\right)\right) = 2 \text{ bits.}$$

Looking at probabilities instead of bits, each position or base pair in the motif multiplies the probability of occurrence by 1/4 (for equal base frequencies). In other words, this model simply multiplies the probability of occurrence of each

specified base/base pair in the motif. For example, this method would say that the motif ACCGUA has the probability of occurrence

$$\text{Prob}(\texttt{ACCGUA}) = \text{Prob}(\texttt{A}) \times \text{Prob}(\texttt{C}) \times \text{Prob}(\texttt{C}) \times \text{Prob}(\texttt{G}) \times \text{Prob}(\texttt{U}). \times \text{Prob}(\texttt{A})$$

The advantages of this method are in its simplicity and fast calculation. Unfortunately, as I showed in [18], the method is also frequently incorrect by several orders of magnitude. Part of the reason for this is that it does not take into account the modularity of motifs or the length of the random sequences, both of which have significant effects on the probability of a motif occurring. Because of this, I do not recommend using the information content approximation.

2.2.3 Poisson Approximation

Another method which is nearly as simple and fast as the information content approximation is the Poisson approximation [37]. Here, we first calculate the probability p of the finding the motif M in a single random sequence that is the same length as the motif. However, we then calculate the number of ways that the motif could appear in the random sequence. For example, a motif might start at the first base in a random sequence, or the second, etc. This calculation takes into account both the length of the random sequence as well as the modularity of the motif. Finally, using the Poisson formula to calculate the probability of zero occurrences in the random sequence and subtracting from 1:

$$\text{Prob}\,(M) = 1 - e^{-p \cdot n},$$

where n is the number of ways the motif can appear in the random sequence.

This approximation assumes that each match of the motif is independent and that matches are very unlikely. These assumptions may be violated for highly-modular motifs since the modularity increases the probability of occurrence. However, as I showed in [18], the Poisson approximation is very close to the exact probability over a wide range of conditions. Furthermore, [17] showed that, for real biological motifs of interest, the Poisson approximation has an average error of about 0.076% for motifs with probability in a range of more than 50 orders of magnitude. Despite these findings, the weakness of the method is that, in general, it is a point estimate with an unknown amount of error and for complex motifs where exact calculations are not feasible, its error is not easily quantifiable.

2.2.4 Upper Bound Approximation

For a random sequence of length l, let W be the number of occurrences of the motif M in the string. Then, $\text{Prob}\,(W \geq 1)$ is the probability of finding M in the random string at least once. We can find an upper bound for this exact probability by making use of Markov's inequality [10], which says that

$$\text{Prob}\,(W \geq 1) \leq \text{E}(W)$$

where $\text{E}(W)$ is the expected value of W. Now, let n be the number of possible ways that M could occur in the random string. For each of these possible ways, let Y_i be equal to 1 if M occurs in this way and 0 otherwise. Then, $W =$

$\sum_{i=1}^{n} Y_i$. Therefore, $E(W) = E(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} E(Y_i) = n \cdot p$, where p is just the probability of the pattern M itself occurring. Therefore, we can calculate an upper bound on the exact probability. Furthermore, by taking the Poisson approximation, $1 - e^{-E(W)}$, and expanding it into a Taylor series, we get

$$\text{Poisson approximation} = 1 - e^{-\mathrm{E}(W)} = \mathrm{E}(W) - \frac{\mathrm{E}(W)^2}{2!} + \frac{\mathrm{E}(W)^3}{3!} - \ldots \leq \mathrm{E}(W)$$

In particular, the upper bound on the exact probability is also an upper bound on the Poisson approximation. Further, we see that the difference between the upper bound and Poisson approximations is of order $\mathrm{E}(W)^2$, which is negligible for small probabilities.

This upper bound approximation is similar to the Poisson approximation in both its speed and simplicity and also provides the additional advantage of being a guaranteed upper bound on the true probability. In [17], I demonstrated that the upper bound approximation agrees very well for a wide range of motifs similar to actual biological motifs. Specifically, I showed that for a motif with an overall probability less than 0.001 and with the probability of any individual module being less than 0.01, the upper bound approximation gives an average error of less than 1%. Furthermore, these conditions can be tested without calculating the exact probability by directly calculating the upper bound probability; if the calculated upper bound approximation satisfies these given constraints (which account for a very wide range of biological motifs), then it can be concluded that the probability is likely very accurate. This stands in contrast to the Poisson approximation, where even if the Poisson probability satisfied the probability conditions, it cannot necessarily be inferred that the true probability does as well.

2.3 Approximation with Confidence Bounds

To deal with the many of the problems inherent in the before-mentioned methods, I have developed a novel technique which makes used of finite automata in a way similar to that of the exact calculation. However, rather than requiring all possible automata to be considered, I use statistical techniques in order to significantly reduce number of calculations required - even for very complex motifs - and instead provide asymptotic confidence bounds on the true probability.

First, we can observe that most of the RNA motifs of interest have probabilities close to zero. Because of this, the probability of several of the non-degenerate representative motifs occurring in the same random sequence - $\text{Prob}(\cap_{i=1,\ldots,k} M_i)$ where $k >> 1$ - is essentially zero relative to the probability of only one or two occurrences. In other words, truncating the entire inclusion-exclusion series (Equation 2.1) at some point is reasonable. By doing this, bounds on the true probability can be given using Bonferroni's inequality [10]:

$$\sum_{k=1}^{2D} M_k \leq \text{Prob(motif)} \leq \sum_{k=1}^{2D-1} M_k, \text{for any } D.$$

Thus, by computing the first $2D$ sums, bounds can be placed around the actual probability. However, because there is no guarantee that the i^{th} sum will provide a tighter bound than the $(i-2)^{th}$, we instead use the inequality

$$\max_{d=1,\ldots,D} \sum_{k=1}^{2d} M_k \leq \text{Prob(motif)} \leq \min_{d=1,\ldots,D} \sum_{k=1}^{2d-1} M_k.$$

By limiting the number of terms that we calculate in the inclusion-exclusion series, we are able to significantly reduce the number of calculations performed. However, for complex motifs the computation necessary is still prohibitive. To deal with this, we use Monte Carlo methods to approximate the value of the sums. For each sum, we are in essence trying to calculate the average of all the values in the sum, since this can then be scaled by the number of terms in the sum to get the actual sum. Then, because of the central limit theorem [10], the distribution of the average of n random samples approaches a normal distribution as $n \to \infty$. For large n, then, confidence intervals can be placed around each sum, naturally leading to asymptotic confidence bounds for the true probability Prob(motif) (after appropriately scaling the value of α for the bounds of each individual sum, as described in [18]).

The advantage of this method is that it provides a quantitative estimate of the error of the approximation, something that none of the previous techniques provide. Also, even though this method is significantly more computationally expensive than other approximations, it is feasible for complex motifs and scales much better than the exact calculation (Figure 2.2). Note that I use the word *asymptotic* when referring to the confidence bounds because the intervals rely on the assumption of a normal distribution, which occurs only asymptotically as $n \to \infty$.

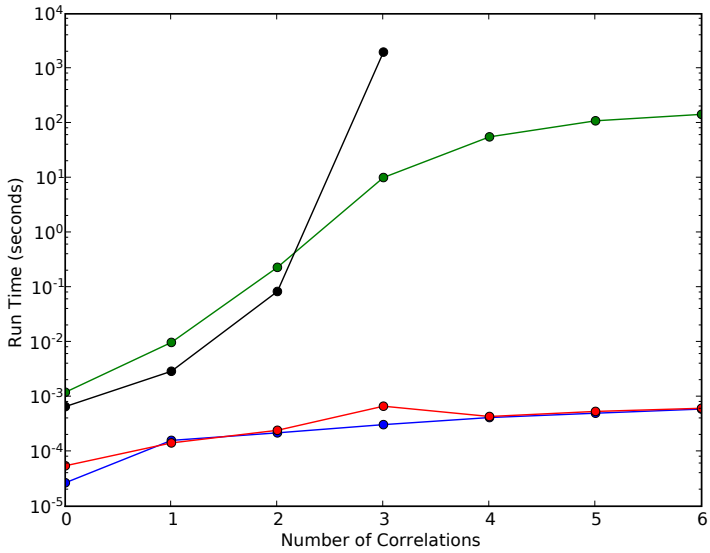

Figure 2.2: Computation time for simple motifs as the number of degenerate correlations increases. Both the information content method (blue) and the Poisson approximation (red) scale very well with increased motif complexity. In contrast, the exact calculation (black) grows fast and the calculation is infeasible for more than three correlations. The asymptotic confidence bound method (green) scales much better than the exact method and remains computationally feasible for complex motifs.

2.3.1 Determining the normality of the distribution

The confidence bounds we are using are asymptotic in the sense that they will not be exact unless the distribution over the number of times a motif occurs in a random string is normal, which only happens in the limit as $n \to \infty$. However, it remains an issue to determine what value for n should be used so that the sampling distribution is nearly normal. To do this, I used the following statistical

hypotheses: H_0: the distribution is normal, H_1: the distribution is not normal. By taking an increasing number of samples from the distribution and testing for normality, it is possible to determine a rough cutoff where the distribution is very nearly normal. Our results indicate that 50 samples were more than enough to ensure near-normality for small, simple motifs while 200 samples were sufficient for large, complex motifs. For normality testing, the omnibus test of D'Agostino and Pearson [8] was used.

2.3.2 Extension to Markovian backgrounds

One assumption that all these methods have made is that the probability of occurrence of any base at a position in a random sequence is independent and identically distributed (i.i.d.). In other words, $\mathrm{P}\left(X_i = x | X_{i-1}, X_{i-2}, \ldots, X_1\right) = \mathrm{P}\left(X_i = x\right)$. Although this model is appropriate for SELEX experiments, where RNAs are actually constructed in this way, it is clearly a simplifying assumption for biological RNAs in the body, where the molecules were created under complex circumstances that are far from i.i.d (for example, it is known that different regions of DNA have significantly different amounts of the bases G and C, known as the regions GC content [12]). We can relax these assumptions, however, by using a Markov model, where we say $\mathrm{P}\left(X_i = x | X_{i-1}, X_{i-2}, \ldots, X_n\right) = \mathrm{P}\left(X_i = x | X_{i-1}, X_{i-2}, \ldots, X_{i-k}\right)$. This is known as a k^{th} order Markov model, and although it is unlikely that actual RNAs were created with such a Markov model, it imposes a less-stringent assumption than does the i.i.d. model.

All the previously-discussed methods make the assumption of an i.i.d. se-

quence. One advantage, however, of using DFAs to compute the probability of occurrence is that they can be easily extended from an i.i.d. model - which is just a 0^{th} order Markov model - to higher order models. This is accomplished by considering the product of an Aho-Corasick automaton with a de Bruijn automaton of order k [23, 29, 30].

CHAPTER 3

Application to the Analysis of RNA Motifs

Using the previously-described probabilistic models, it is possible to calculate the probability that an RNA motif occurs in a random sequence of nucleotide bases. I now use these models in the analysis of actual biological RNA motifs. A set of about 30 motifs which were artificially selected from SELEX experiments were taken from the literature, covering a range of functions. For each motif, I calculated the probability of the sequence elements occurring in a random string at 5% intervals in composition space such that each base (A,C,G,U) had at least a 5% composition. Plotting this allows us to see where in composition space the motif is most likely to occur (Figure 3.3). However, having just the sequence elements of a motif does not guarantee that it will fold correctly. To take this into account, I used the RNAFold program from the Vienna package [15] to estimate the probability that the motif will fold correctly given that the sequence elements are present. Multiplying these two probabilities together gives the joint probability of the motif sequence occurring *and* folding correctly: Prob(Sequence)·Prob(Folding|Sequence) = Prob(Folding&Sequence). This could be taken even further by estimating Prob(Function|Folding&Sequence), but this would require lab expereriments that were not covered in this work.

Once the probabilities were calculated for all motifs, comparing the distributions of different motifs would give an indication of how the distributions differ by motif. However, because the probabilities were calculated at the same locations in composition space for every motif, overlaying the plots would be uninformative and comparing them side-by-side would be difficult for more than a few motifs. Instead, I fit a multivariate Gaussian distribution to the distribution of each motif. Then, by plotting the one-standard-deviation ellipsoids, I was able to easily compare the high-probability regions of every motif at once. Of course, the use of a Gaussian distribution is not necessarily well-founded because the underlying data are almost surely non-normal. However, the ellipsoids were used for visualization and for this application they provided a convenient way to compare different motif distributions (Figure 3.1).

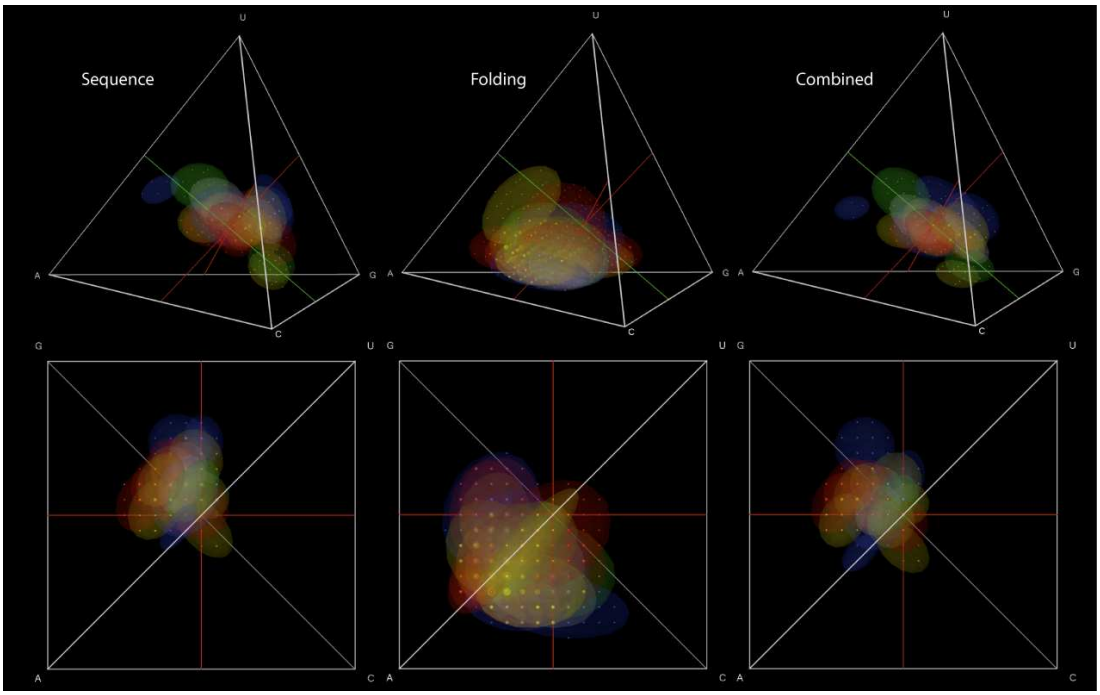

Figure 3.1: Distribution of one-standard-deviation ellipsoids for Sequence, Folding and Sequence&Folding

Although examination of these artificially-selected motifs may give interesting conclusions, even more information can come from a comparison to naturally-occurring motifs. For this comparison, sequences from Rfam [13] were used. In particular, the aptamers (riboswitches) and ribozymes from Rfam were compared to our artificially selected motifs because only these RNAs are themselves functional. In fact, the other classes of RNAs from Rfam - splisosomal, snRNA, and miRNA - did not follow the same distribution as the ribozymes and riboswitches. (Figure 3.2).

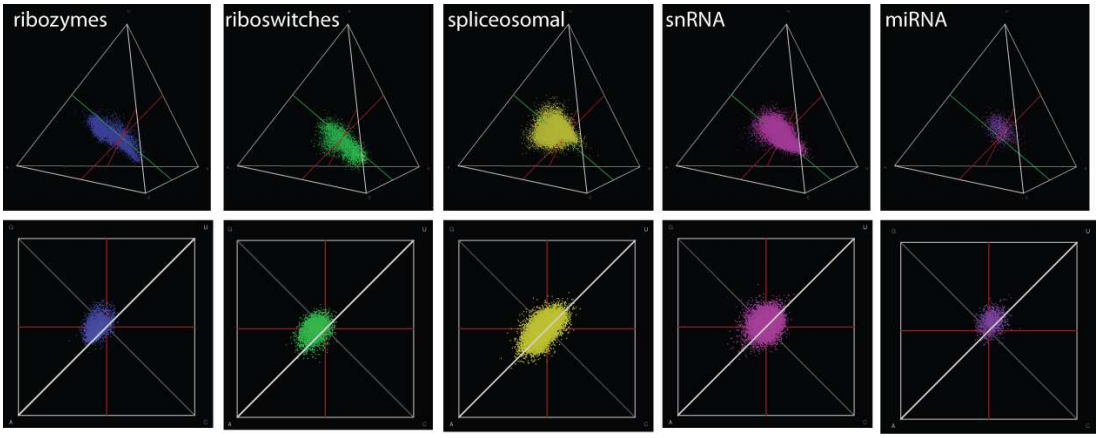

Figure 3.2: Distribution of sequences taken from Rfam, by function.

Figure 3.3 shows the comparison between the distributions of artificially-selected motifs and of naturally-occurring aptamers and ribozymes. The overlap of the two distributions is striking. In particular, both distributions exhibit a statistically significant bias toward purines (shown more clearly in Figure 3.1 for artificially selected motifs) as well as a large spread along the GC axis (the axis with the same proportion of the bases G and C) as compared to the two axes orthogonal to it. Because these similarities exist between artificially-selected and naturally -ccurring sequences, we can conclude that these features are not a result of intermolecular interactions because the artificially selected motifs have

never been inside a living cell. Instead, these biases must be the result of general requirements for functional RNAs. Examining Figure 3.1 sheds even more light on the situation because it shows that the purine bias is the result of an interaction between the sequence probability - which biases toward G - and the folding probability - which biases toward A.

Another important conclusion can be seen by examining the colors in Figure 3.1. The colors in the figure correspond to different functions of the RNAs. It is evident from the figure - and is also a statistically significant finding - that RNAs with different functions do not occupy distinct regions of the composition space. Instead, all the functions overlap within the same distribution and there is no significant difference between the spaces that they occupy. This has important implications for evolution; it supports the idea that it is possible for RNAs to evolve into a new function without radical changes. Because all functions share a similar region of composition space, an RNA could perform one function right up until a single mutation that would cause the RNA to perform a new function rather than having to undergo significant changes between the two functions.

Furthermore, the overlap of functions as well as the purine bias have implications for designing successful SELEX experiments. In order to create a large number of functional sequences, it is best to bias the composition of the sequences toward the purines, rather than using equal base frequencies or trying to tune the composition toward that of RNAs of the desired function.

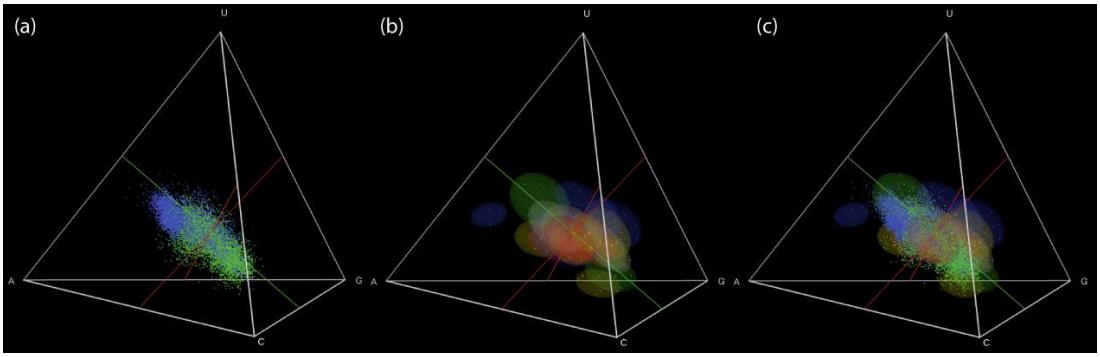

Figure 3.3: Comparison of distributions of (a) natural aptamers (green) and ribozymes (blue) and (b) artificially selected motifs. The two distributions are superimposed in (c).

CHAPTER 4

Searching for Patterns in Sequence Data

Until now, I have looked at information concerning pre-specified RNA motifs. However, for much of the available biological sequence data, there may not be specific motifs that have been identified already, and this problem will only worsen with the exponential growth in sequences (GenBank [2], for example, currently has about 100 million sequences and has been doubling every 18 months). In order to extract information from these sequences, it is important to have techniques that do not depend on the availability of motifs but can be used on raw sequence data. Toward this goal, I present a framework of methods that can be applied to this type of data.

4.0.3 Background

As a first application of the technique I will describe, I look at data associated with the transcription factor p53. Transcription factors (TF) are proteins that bind to DNA and affect transcription (the process of creating RNA from the DNA blueprint). p53 is a well-studied transcription factor because of its known

effects on tumor suppression.

In the analysis, I used two data sets:

1. Horvath, et al. 2007 [16] provides 81 p53 TF binding sites (TFBS) and the associated gene functions (i.e., apoptosis, cell cycle, etc.).
2. Riley, et al. 2008 [32] provides 157 p53 TFBS and whether they are activators or repressors.

In analyzing these data sets, our primary goal is to see if there are specific factors that determine whether p53 acts as an activator/repressor or why p53 binds to genes with different functions, depending on which data set is used. I examine two different hypotheses. The first is that the *sequence* of the binding sites themselves affect p53's function, while on the other hand it may be the *context* of the p53 molecule - its interaction with other nearby transcription factors - that is the determining factor.

4.0.4 Effect of the Sequence

Our first hypothesis is that the sequence of the p53 binding sites themselves are the determining factor for the differences between activator/repressor or gene function. In other words, this supposes that sequences that are similar to each other will also be similar in their function. In order to determine whether two sequences are similar, two different distance metric are used. The first metric is

the Levenshtein distance - also known as the edit distance [22]. The Levenshtein distance is defined for any two strings as the number of *insertions*, *deletions* and *substitutions* that are required to transform one string into the other; if the two strings are of the same length this is just the commonly-used Hamming distance. Alternatively, we can use known information about DNA to try to improve our distance metric. It is known that changes within the groups {A,G} and {C,T} - called transitions - are more common than changes across the two groups - called transversions [28]. Thus, we can define a new metric which I will call the *alternative distance*, which is defined as the Hamming distance between two equal-length strings except that transitions are counted as 1 edit while transversions are counted as 2. The alternative distance requires that the two strings be the same length because the transitions/transversions are considered between the same base, while for unaligned sequences of varying length it is not possible to determine which bases should be paired.

4.0.5 Visualization of the data sets

As a first approach, the distance matrix for a given sequence alignment can be visually inspected. To do this, I used principal coordinates analysis [3] (PCoA). PCoA is a visualization technique when only the distances between points are known. For example, suppose that we want to visualize how the cities of Denver, Colorado Springs, Miami, Philadelphia are spread out on a map when the only available information is the distance between each pair of cities, as in Table 4.1. PCoA will take this distance matrix and convert the distances into points such that the distances between points corresponds to the values in the supplied

distance matrix. In addition, the points are generated such their axes are ordered by the amount of variance in the data that they explain. In this case, the data can be explained using just two axes (since a map is two-dimensional) and this plot is shown in Figure 4.1. As expected, Denver and Colorado Springs appear much closer to each other than to Miami or Philadelphia.

	Denver	Colorado Springs	Miami	Philadelphia
Denver	0	63	1725	1578
Colorado Springs	63	0	1690	1582
Miami	1725	1690	0	1019
Philadelphia	1578	1582	1019	0

Table 4.1: Distance matrix for several U.S. cities, in miles. Data from [1].

When visualizing sequence data, however, the resulting points are in a space that is typically of a much higher dimension than two or three and it is not easy to view the points directly. Instead, plotting the data on just the first two principal components will give the most information about the data, with successive axes explaining less and less of the variance. After plotting the data in this way, I colored the points according to the function of their corresponding sequences with the hope that points of the same color would tend to cluster together.

To demonstrate the use of PCoA, I downloaded the sequence alignments for the RNAs miR-101 and snoME28S-Cm2645 from Rfam. These two RNAs have very different sequences and so should serve as a good positive control for our PCoA technique. The results from this data set are shown in Figure 4.2. Clearly, these points form two clusters to the extent that almost 95% of the variance in the data can be explained by only a single axis.

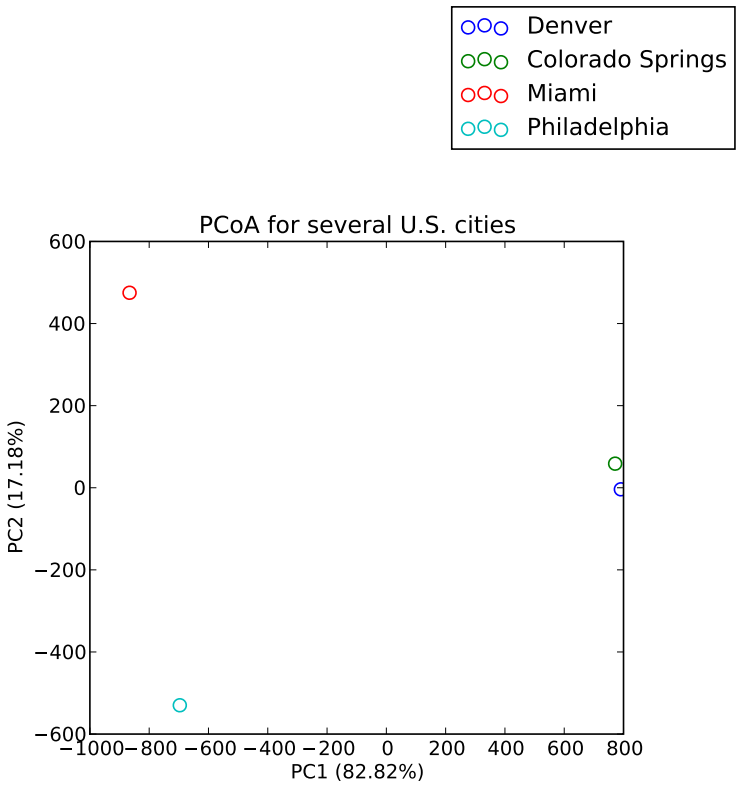

Figure 4.1: PCoA for several U.S. cities.

I then applied PCoA to the Horvath data set and colored the points by gene function to see whether a similar clustering occurred (Figure 4.3). Interestingly, the clustering for the Horvath data set is minimal if existent at all. Using the alternative distance provided similar results (Figure 4.4), although the first axis does explain slightly more of the variance in the data.

The same technique can be applied to the Riley data set (Figure 4.5), where only the Levenshtein distance could be used because not all binding sites were of identical length. Again, very little clustering seems to have occurred.

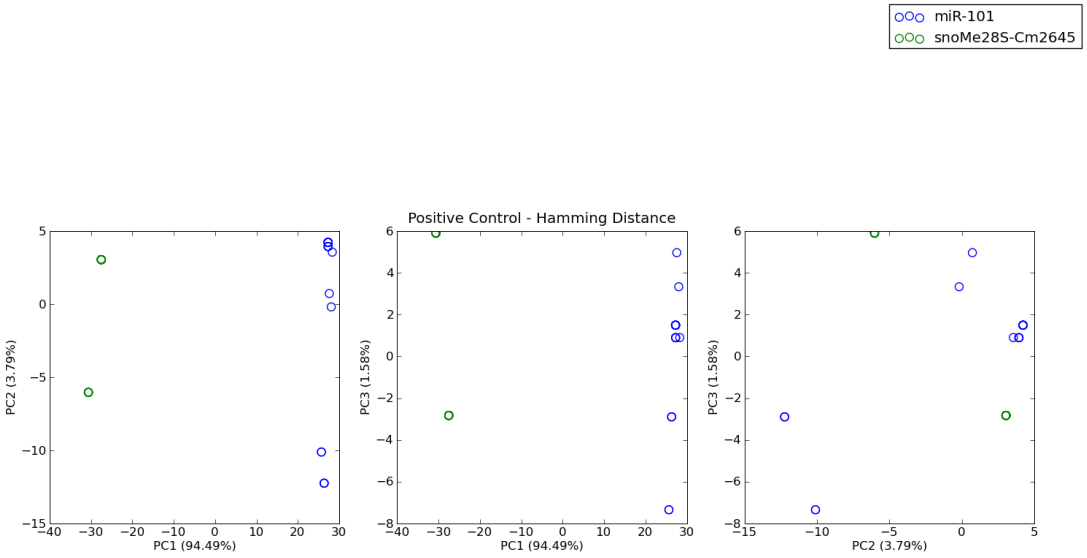

Figure 4.2: PCoA for positive control.

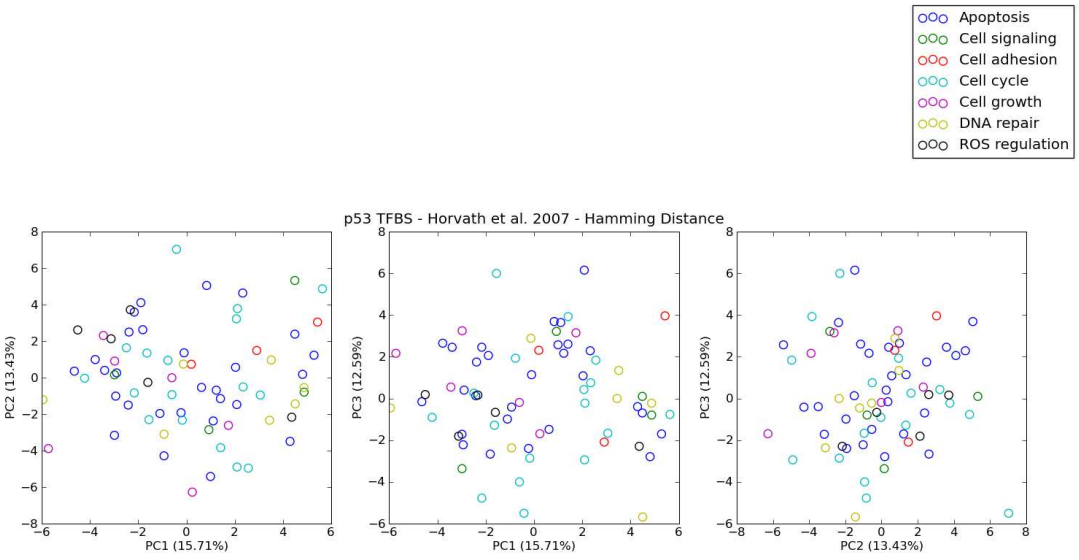

Figure 4.3: PCoA for Horvath data set using Hamming distance.

4.0.6 Supervised classification

Even though visual inspection seems to show no clustering in these data sets, it might be possible that there is *some* significant clustering but that it might be

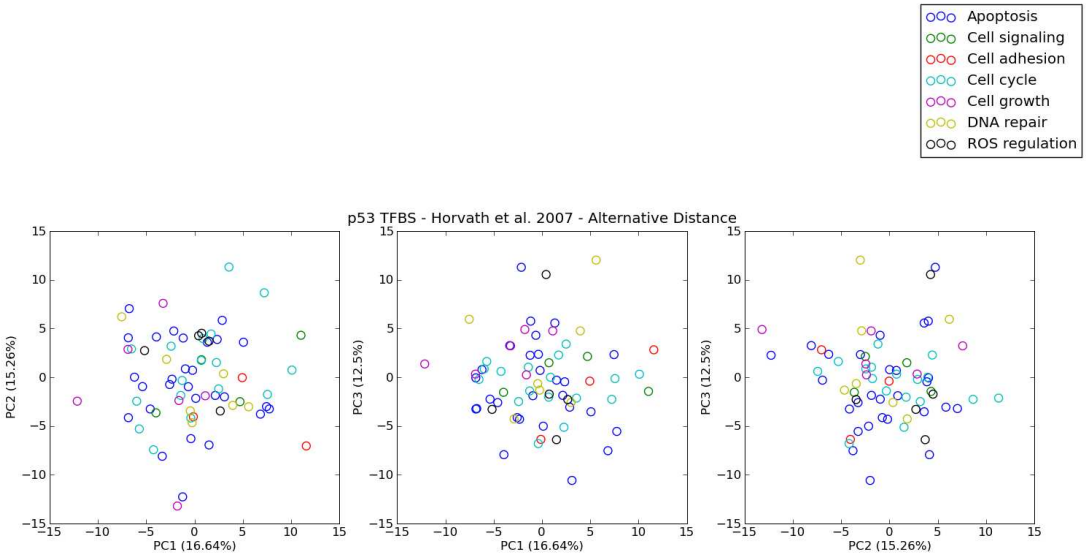

Figure 4.4: PCoA for Horvath data set using alternative distance.

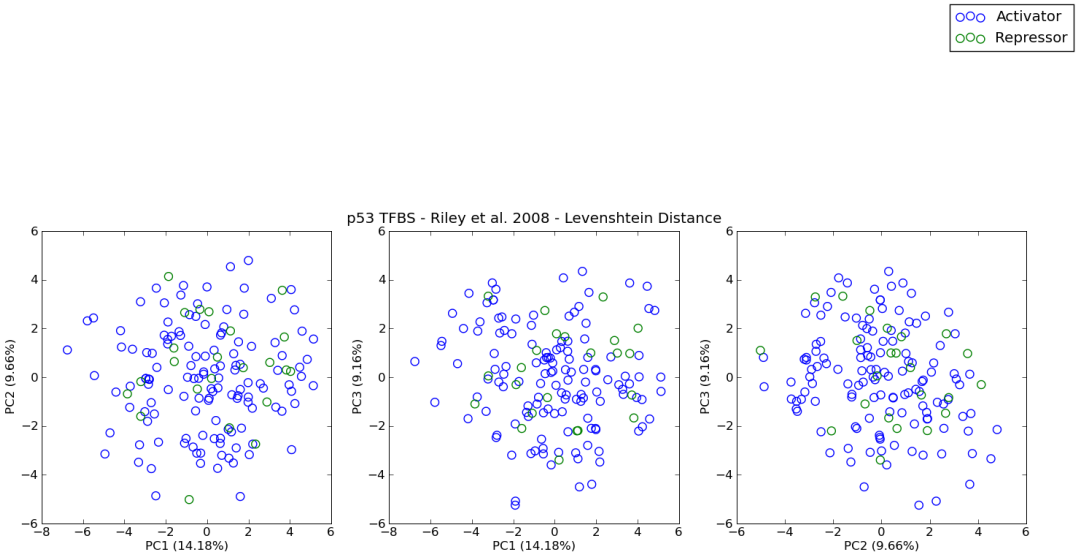

Figure 4.5: PCoA for Riley data set using Levenshtein distance.

too subtle to see visually or might require a view of the entire high-dimensional space rather than just the first few principal components. For this, I used a super-

vised classification technique. Given a set of data points along with the different classes (i.e., activator/repressor or gene function) for each point, a supervised classification algorithm attempts to find a boundary that separates the different classes as well as possible. For this thesis, I used random forests [6], a technique that is based on decision trees. Random forests were used because they are good classifiers, are fast, automatically generate an unbiased estimate of classification error on unseen data, and also allow the use of sequence data directly.

To demonstrate the use of random forests, I return to the miR-101 and snoME28S-Cm2645 sequences. By feeding the points generate by PCoA, as well as whether each point corresponded to miR-101 or snoME28S-Cm2645, into the random forest, the confusion matrix of Table 4.2 was produced, which shows how the points in each class were classified. The out-of-bag error estimate (OOB error; an estimate of future error on unseen data) was 5.13% when I would expect an error rate of 48.72% by chance (the least error I would expect were there no correlation between points and their corresponding classes; calculated as simply choosing the class with the largest number of points for every classification). As before, this positive control demonstrates that the use of these techniques does provide useful information if any patterns exist.

	miR-101	snoMe28S-Cm2645	Error
miR-101	18	2	10%
snoMe28S-Cm2645	0	19	0%

Table 4.2: Confusion matrix for positive control using the Hamming distance for random forest classification.

The same technique was first applied to the Horvath data set and the resulting confusion matrix is shown in Table 4.3. Here, the OOB error rate is 50%, while

	Apop.	Cell Ad.	Cell Cy.	Cell Gr.	Cell Sig.	DNA Rep.	ROS Reg.	Error
Apop.	26	0	2	0	0	0	0	7.14%
Cell Ad.	2	0	1	0	0	0	0	100%
Cell Cy.	8	0	9	0	0	0	0	47.06%
Cell Gr.	2	0	4	0	0	0	0	100%
Cell Sig.	3	0	1	0	0	0	0	100%
DNA Rep.	3	0	4	0	0	0	0	100%
ROS Reg.	3	0	2	0	0	0	0	100%

Table 4.3: Confusion matrix for Horvath data set (7 classes) using the Hamming distance for random forest classification. Class names are abbreviated for space.

by chance I would expect 60%, so random forests seem to be able to extract at least a little information from the data points. It is also notable that the only two classes that have less than 100% error are the apoptosis and cell cycle classes, possibly because they are the largest two classes. If I look at only the points in these two classes, I get the results in Table 4.4. The OOB error rate for this data

	Apoptosis	Cell Cycle	Error
Apoptosis	25	3	10.71%
Cell Cycle	9	8	52.94%

Table 4.4: Confusion matrix for Horvath data set (2 classes) using the Hamming distance for random forest classification.

set is 26.67%, and I would expect 37.78% by chance. Again, there does seem to be some effect that random forests are able to discern, although the effect might have been too subtle for us to see visually with PCoA. Using the alternative distance did not result in any better classification.

The same analysis can be applied to the Riley data set, giving the results shown in Table 4.5. The OOB error for this data set is only 3.67%, with chance

	Activator	Repressor	Error
Activator	131	1	0.76%
Repressor	5	18	21.74%

Table 4.5: Confusion matrix for Riley data set using the Levenshtein distance for random forest classification.

being 14.84% because the data set is so heavily biased toward activators. Interestingly, this is the only random forest analysis that has yielded less than 50% error for every class. This might indicate that the sequence does have some effect at separating data points, although a larger data set with more repressors would be needed to verify this finding.

One advantage of using random forests for classification is that they allow the use of non-numeric data. Instead of converting the sequences into a distance matrix and then using PCoA to create a set of points that can be input to the random forest, the sequences could stand for themselves. This does require that the sequences be the same length, though, and so in order to avoid dealing with gaps in the alignment of the Horvath data set, I looked only at the Riley data set. Using the sequences themselves as input to the random forest allows us to see

whether there are any differences in the actual sequences between classes rather than just in the distances. Table 4.6 shows the confusion matrix for the positive control data set using the sequence data. The random forest is able to separate

	miR-101	snoMe28S-Cm2645	Error
miR-101	20	0	0%
snoMe28S-Cm2645	0	19	0%

Table 4.6: Confusion matrix for positive control using the Hamming distance for random forest classification, using the sequence data directly.

the two classes perfectly, indicating that there is some intrinsic difference in the sequences of the two classes.

Applying this analysis directly to the Horvath data set gives the confusion matrix shown in Table 4.7. Unfortunately, using the sequence data performs worse that using the distance data, so there is likely no difference in the sequences themselves between data sets. Similar results were seen for the 2-class Horvath data set (Table 4.8).

4.0.7 Detecting covariation in sequences

Whether or not separating the classes based on their sequences yielded any patterns, there might be additional information due to base pairing in the sequences that the previous methods were unable to pick up. For example, consider the following alignment:

	Apop.	Cell Ad.	Cell Cy.	Cell Gr.	Cell Sig.	DNA Rep.	ROS Reg.	Error
Apop.	19	0	7	1	0	0	0	32.14%
Cell Ad.	1	0	2	0	0	0	0	100%
Cell Cy.	14	0	1	0	1	0	1	94.12%
Cell Gr.	5	0	1	0	0	0	0	100%
Cell Sig.	1	0	3	0	0	0	0	100%
DNA Rep.	5	0	2	0	0	0	0	100%
ROS Reg.	5	0	0	0	0	0	0	100%

Table 4.7: Confusion matrix for Horvath data set (7 classes) using the Hamming distance for random forest classification, using the sequence data directly. Class names are abbreviated for space.

	Apoptosis	Cell Cycle	Error
Apoptosis	21	8	27.59%
Cell Cycle	15	2	88.24%

Table 4.8: Confusion matrix for Horvath data set (2 classes) using the Hamming distance for random forest classification, using the sequence data directly.

AGAGGUA

UCAUGUA

UCGCCCG

CCUGAAU

GUCAUGA

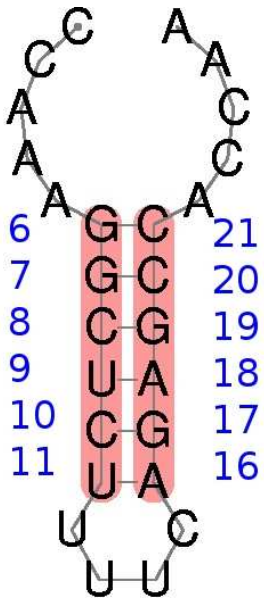

Figure 4.6: Secondary structure of the histone3 RNA. Adapted from Rfam.

At first glance, it might appear that the underlined 3^{rd} and 6^{th} bases are unpredictable; they could be any base. However, when viewed together, it becomes evident that the two bases are Watson-Crick pairs; an A always accompanies a U, and likewise for C and G. In fact, it might be the case that it is not important what the actual bases are, but instead it is important that they are able to pair together.

To detect covarying positions like this, I looked at the *mutual information* between the two bases [3, 7]. Formally, the mutual information is defined as $M(X,Y) = H(X) + H(Y) - H(X,Y)$, where $H(X)$ and $H(Y)$ is the entropy of the positions X and Y, respectively, and $H(X,Y)$ is the joint entropy of the two. More intuitively, the mutual information tells how predictable the value of X is given the value of Y, and vice versa. Here, I actually use the mutual information normalized by the joint entropy of the positions ($NM(X,Y) =$

$M(X,Y)/H(X,Y)$) which controls for different rates of coevolution [7, 26].

As an example of what this technique shows us, I looked at Histone3 [9], an RNA taken from Rfam. It's secondary structure, depicted in Figure 4.6, shows that the molecule has known base pairing. In order to determine any high levels of mutual information, I first calculated the mutual information between every position in the alignment and then plotted this as a heatmap, shown in Figure 4.7

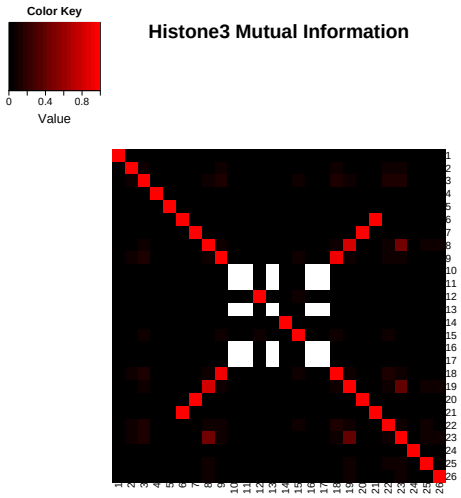

Figure 4.7: Mutual information heatmap for histone3.

The first noticeable feature is the high mutual information along the main diagonal. This is actually just because the mutual information of any base with itself is 1. The white squares in the center of the figure are positions for which the normalized mutual information is undefined because the positions' bases did not vary at all and so had a join entropy of 0. However, there is also an interesting diagonal of high mutual information orthogonal to the main diagonal. The base pairs - (6,21), (7,20), (8,19), and (9,18) - are in fact known base pairing positions

as was seen in the secondary structure of Histone3. The remaining base pairs are undefined for their normalized mutual information, but if I had a larger alignment we would probably see a high level of mutual information between them as well.

The mutual information heatmap for the entire Horvath data set is shown in Figure 4.8. Although there are no clear patterns of mutual information, the hope

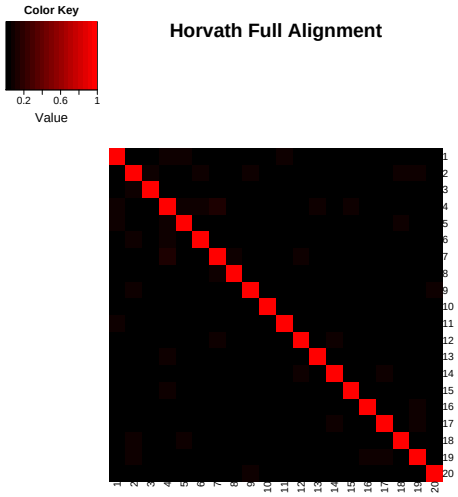

Figure 4.8: Mutual infomation heatmap for Horvath - full alignment.

is that looking at the alignments by gene function would vary based on function. Unfortunately, as Figures 4.9 and 4.10 show, there are no regions of high mutual information within the different classes of binding site either. Looking for mutual informaton in the Riley data set had similar results (Figures 4.11, 4.12 and 4.13).

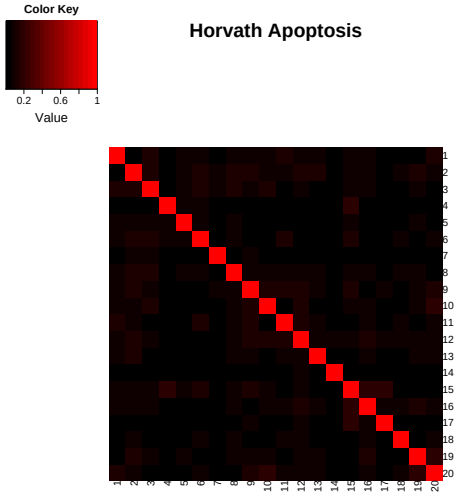

Figure 4.9: Mutual infomation heatmap for Horvath - apoptosis sites.

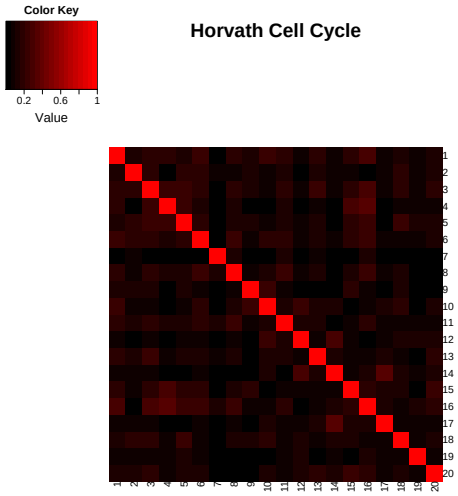

Figure 4.10: Mutual infomation heatmap for Horvath - cell cycle sites.

4.0.8 Effect of Context

It is possible that there is very little difference in the sequences for different binding sites. Instead, it could be possible that the p53 protein interacts with

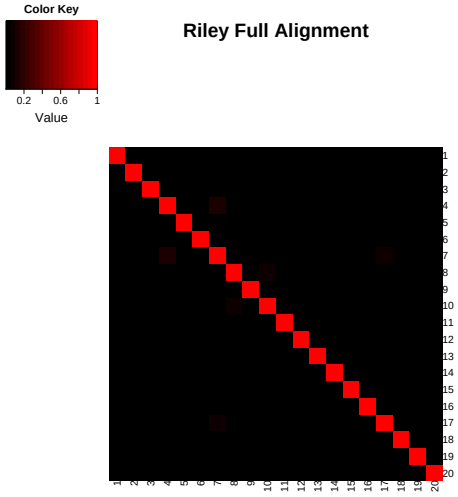

Figure 4.11: Mutual infomation heatmap for Riley - full alignment.

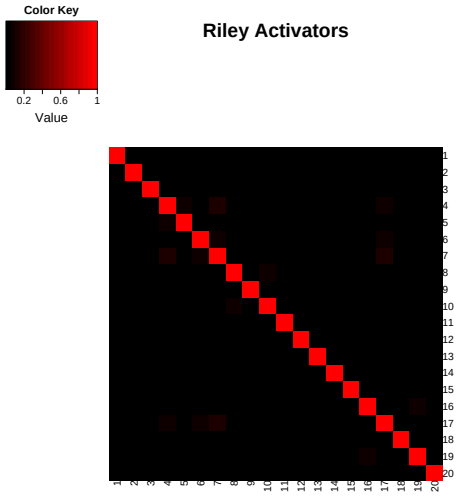

Figure 4.12: Mutual infomation heatmap for Riley - activators.

nearby transcription factors which causes it to function differently. In testing this hypothesis I used only the Riley data set because it was the more extensive of the two and because it provided the location of the binding sites on the genes so that the 5000-nucleotide sequence upstream from the promoter (where the

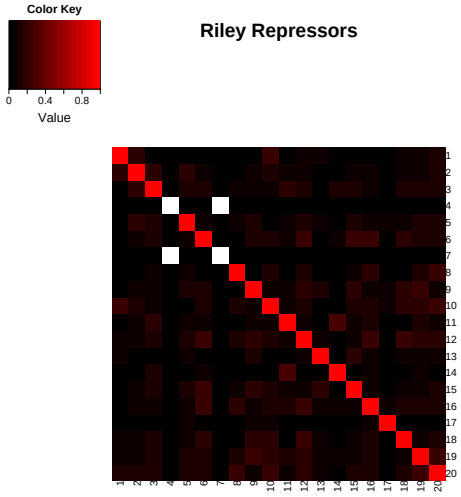

Figure 4.13: Mutual infomation heatmap for Riley - repressors.

p53 binding site lies) could be gathered (sequences were taken from the UCSC Genome Browser database [11]). Also, a list of transcription factors that might interact with p53 were collected from TRANSFAC [27]. The list included the factors C/EBP, E2A, MyoD, E2F1, E2F2, SP1, IRF1, STAT1, STAT3, AP2, VDP, HIF-1A, BRCA-1, FOXO-1, FOXO-3, FOXO-4, P3000, TBP, YY1, CTF2, NFκb, NFY and AR.

The following procedure was used to determine the effect of context on p53. First, the probability weight matrices (PWM) for each TF were taken from TRANSFAC (or inferred from the consensus sequence if it was not available). The probability weight matrices are matrices that specify, for each position in a binding site, what the probability is of that position being either A, C, G or T (the PWM for STAT1 is shown in Table 4.9). The PWM is then converted to a log-odds matrix by taking, for each element in the PWM, $log(p/f)$, where p is the probability given by the PWM and f is the frequency of the base in the

background composition (i.e., the frequencies of each base in the entire length of each promoter I examined). Then, given any sequence of bases the same length as the binding site taken from the promoter sequence, the log-odds matrix will give a score that tells how likely it is that the site is a binding site. However,

Position/Base	A	C	G	T
1	0.23	0.33	0.35	0.10
2	0.38	0.17	0.17	0.27
3	0.15	0.13	0.12	0.60
4	0.00	0.00	0.04	0.96
5	0.02	0.00	0.02	0.96
6	0.29	0.67	0.04	0.00
7	0.08	0.56	0.15	0.21
8	0.19	0.31	0.31	0.19

Table 4.9: Probability weight matrix for STAT1

given this score, there is still an issue of determining which are true binding sites; I therefore needed a cutoff score. In order to determine this, I looked at p53, for which I already had a large number of verified binding sites. Figure 4.14 shows histograms of the scores of true p53 binding sites and randomly-generated sites from the same background composition. Based on this information, I chose to use a cutoff score of 8 so that about 99.6% of the probability mass to the right of the threshold constituted true binding sites. Based on this, I could determine where binding sites for other transcription factors were and their distance to the verified p53 site.

Constructing features which could be used in a supervised learning algorithm

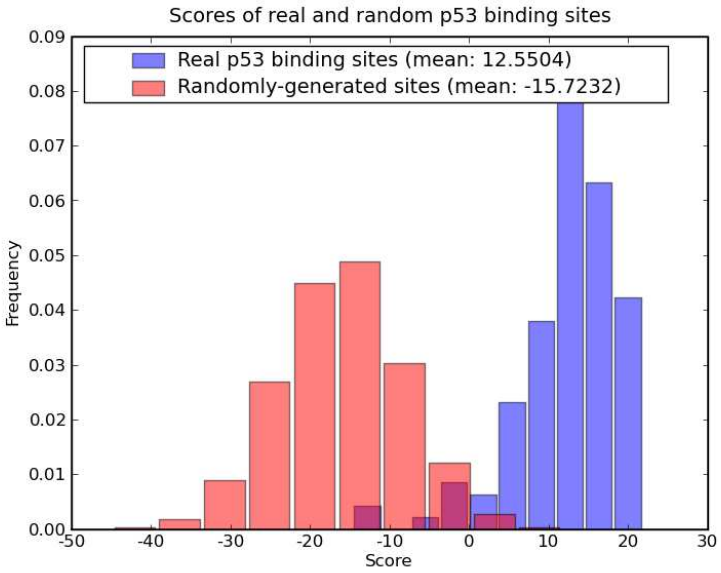

Figure 4.14: Distributions of p53 log-odds matrix scores for verified binding sites and random sequences.

for this data was tricky. In order to obtain a vector representation for each p53 site of the same length, the distance of the top-scoring site that exceeded the threshold of 8 was used as an element in the vector, with TFs that had no sites above the threshold set to 5000 (the length of the promoter sequence used). Along with the classes as labels, the vectors were input into a random forest algorithm and the resulting OOB error rate was 13.16%, while by chance I would expect 14%. The difference here is very slight, though, and increasing to the top 2 or 3 binding sites for each TF or changing the representation of features did not result in any further improvement.

CHAPTER 5

Discussion

In this thesis, I have presented novel methods for analyzing nucleic acids in biology. The first of these techniques involved computing the probability of occurrence of an RNA motif. Our analysis of these methods and contribution of a new one has provided a way for researchers to choose an appropriate technique that balances efficiency and accuracy; while the information content method is typically inaccurate, both the Poisson and upper bound approximations are accurate for a wide range of different motifs and a new method using confidence intervals provides a quantitative estimate of error in the approximation at the cost of higher computational complexity. Also, by using the upper bound approximation it is possible to determine whether any motif meets the criteria under which I've determined these methods to be accurate. This is an important point: given any motif, if the upper bound probabilities give results that meet the criteria outlined in Section 2.2.4 then the probability that was just calculated is likely a very accurate approximation, and if a more rigorous assurance of accuracy is needed then our confidence interval method will provide it.

There are, however limitations to these techniques which could be addressed

in future research. First of all, most of these techniques assume the distribution of nucleotide bases is i.i.d., while actual biological RNA motifs have surely not evolved in this way. Instead, naturally-occurring RNAs have been created by a complex network of interactions which would be infeasible to model in its totality. However, our confidence interval method partially overcomes this limitation by generalizing to a Markov background and future applications of this model might lead to more accurate results on biological motifs. Even the Markov model is limited, though, and is not fully representative of biological motifs. Modeling sequences using a non-homogeneous Markov model - or other more complex models - might be more appropriate, at the expense of higher complexity.

Despite the simplifying assumptions, applying these techniques to both artificially- and naturally-selected motifs has resulted in conclusions that make a significant contribution to our understanding of how RNAs have evolved. First of all, they have helped explain compositional biases observed in both artificial and natural motifs. Many compositional biases occur in biological processes, such as an over-representation of the bases G and C in some regions of DNA or the purine bias observed in naturally-occuring RNAs [33]. Even when biases such as these are fairly ubiquitous, explaining them is not always simple. In the results presented in this thesis, however, we have taken a step toward understanding the purine bias of many RNAs. Specifically, I have shown that the purine bias of many natural RNAs is also present in artificially-selected sequences and so cannot be explained by intercellular interactions. Instead, the bias must be a consequence of constraints on the RNAs themselves. Furthermore, we discovered that the constraints imposed by having the correct sequence elements for the motif tend to bias the composition in a way that is distinct from the biases imposed by the folding constraints; combined, these two separate requirements favor RNAs with

a purine-heavy composition.

Furthermore, the result that functional motifs do not occupy distinct regions of composition space and have an overall purine bias has direct implications for the design of SELEX experiments, especially since the i.i.d. sequence model is consistent with how SELEX is performed. I have shown that imposing a slight purine bias will increase the probability of finding functional motifs, no matter what the function is. This is also important to our understanding of the origin of life. As complex as life currently is, it was not always so. Indeed, life began as very simple structures, with the first organisms possibly lacking DNA and protein altogether [14]. Instead, they might have used RNA for many different functions. Even in this simple world, however, complex RNAs must have evolved from simpler ones. If RNAs with different functions had very different compositions, then it would have been difficult to evolve new functional groups of RNA because the jump from one function to another might involve a complex set of mutations that drastically change the composition of the RNA. However, I have shown that this is not so; many different RNAs with different functions occupy the same region of composition space. Evolving new RNAs then might have been possible by just a few simple mutations.

As the number of sequences available grows exponentially, the number of motifs and complexity of their representation will also grow and this same analysis could be applied to a larger number and wider variety of motifs in future research. Also, extending the analysis from looking at the probability of a motif having the sequence elements and folding correctly to include whether it is functional could add another layer of information to this analysis. We plan on taking this step by producing sequences in a lab that meet the specifications for sequence and

folding and testing whether or not they are functional.

In this thesis I also presented a framework of machine learning techniques that can help in finding patterns in sequence data. Interestingly, the application of these techniques to p53 resulted in few significant results. This suggests that the interactions between p53 and other biological molecules is a complex, dynamic process that is difficult to model. However, understanding these interactions is nonetheless very important for fields such as cancer research, since p53 has a role in tumor suppression. Also, the techniques presented here could be applied to many other domains of research in molecular biology. As the vast repository of biological data only continues to grow, these methods will also need to be extended to search for different patterns on other datasets.

It is important to realized that the analyses I have presented here are far from final products. Instead, these results build on an enormous amount of research that has been done previously and provide only a higher base from which future researchers can continue. The world of biology has many more discoveries to be made and finding them will require analysis techniques such as those presented here.

References

[1] Travel distance between cities, airports, and countries, April 2009. `http://www.convertunits.com/distance/`.

[2] D. A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L Wheeler. Genbank. *Nucleic Acids Research*, 36:D25–30, 2007.

[3] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, 2007.

[4] K. F. Blount and O. C. Uhlenbeck. The hammerhead ribozyme. *Biochem Soc. Trans.*, 30:1119–1122, 2002.

[5] V. Bourdeau, G. Ferbeyre, M. Pageau, B. Paquin, and R. Cedergren. The distribution of rna motifs in natural sequences. *Nucleic Acids Research*, 27:4457–4467, 1999.

[6] L. Breiman. Random forests. *Machine Learning*, 41:5–32, 2001.

[7] J. G. Caporaso. *Extracting signal from noise in biological data: evaluations and applications of text mining and sequence coevolution*. PhD thesis, University of Colorado, 2009.

[8] R. B. D'Agostino. Simple portable test of normality: Gearys test revisited. *Psychological Bulletin*, 74:128–140, 1970.

[9] Z. Dominski and W. Marzluff. Formation of the 3' end of histone mRNA. *Gene*, 239:1–14, 1999.

[10] R. Durrett. *Probability: Theory and Examples*. Duxbury Press, New York, NY, 2004.

[11] D. Karolchik et al. The ucsu genome browser database: 2008 update. *Nucleic Acids Res.*, 36:D773–9, 2008.

[12] S. Freeman. *Biological Science*. Pearson Custom Publishing, 2008.

[13] P.P. Gardner, J. Daub, J.G. Tate, E.P. Nawrocki, D.L. Kolbe, S. Lindgreen, A.C. Wilkinson, R.D. Finn, S. Griffiths-Jones, and S.R. et al. Eddy. Rfam: updates to the RNA families database. *Nucleic Acids Res.*, 2008.

[14] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.

[15] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Chemical Monthly*, 125:167–188, 1994.

[16] M. M. Horvath, X. Wang, M. A. Resnick, and D. A. Bell. Divergent evolution of human p53 binding sites: Cell cycle versus apoptosis. *PLoS Genetics*, 3:e127, 2007.

[17] R. Kennedy, M. Lladser, Z. Wu, C. Zhang, M. Yarus, H. De Sterck, and R. Knight. Active site specifications and self-organization drive universal compositional biases in naturally and artificially selected RNA sequences. *Manuscript in preparation for submission to Nature*, 2009.

[18] R. Kennedy, M. Lladser, M. Yarus, and R. Knight. Information, probability, and the abundance of the simplest RNA active sites. *Frontiers in Bioscience*, 13:6060–6071, 2008.

[19] U. Laserson, H. Gan, and T. Schlick. Predicting candidate genomic sequences that correspond to synthetic functional RNA motifs. *Nucleic Acids Research*, 33:6057–6069, 2005.

[20] D. Lazarev, I. Puskarz, and R. R. Breaker. Substrate specificity and reaction kinetics of an x-motif ribozyme. *RNA*, 9:688–697, 2003.

[21] M. Legiewicz and M. Yarus. A more complex isoleucine aptamer with a cognate triplet. *J Biol Chem*, 280:19815–19822, 2005.

[22] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

[23] M. Lladser. Minimal markov chain embeddings of pattern problems. *Information Theory and Applications Workshop*, pages 251–255, 2007.

[24] M. Lladser, M. Betterton, and R. Knight. Multiple pattern matching: a markov chain approach. *J Math Biol*, 356:51–92, 2008.

[25] I. Majerfeld and M. Yarus. A diminutive and specific RNA binding site for l-tryptophan. *Nucleic Acids Res.*, 33:5482–5493, 2005.

[26] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 22:4116–4124, 2005.

[27] V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, Chekmenev, D., M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, and E. Wingender. Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34:D108–110, 2006.

[28] M. E. Mulligan. Mutations & mutagenesis, April 2009. `http://www.mun.ca/biochem/courses/3107/Topics/Mutations.html`.

[29] P. Nicodème. Regexpcount, a symbolic package for counting problems on regular expressions and words. *Fundamenta Informaticae*, 56:71–88, 2003.

[30] P. Nicodème, B. Salvy, and P. Flajolet. Motif statistics. *Theoretical Computer Science*, 287:593–617, 2002.

[31] Nomenclature Committee of the International Union of Biochemistry. Nomenclature for incompletely specified bases in nucleic acid sequences. *Recommendations. Eur. J. Biochem.*, 150:1, 2985.

[32] T. Riley, E. Sontag, P. Chen, and A. Levine. Transcriptional control of human p53-regulated genes. *Nat Rev Mol Cell Biol*, 9:402–412, 2008.

[33] E. Schultes, P. T. Hraber, and T. H. LaBean. Global similarities in nucleotide base composition among disparate functional classes of single-stranded rna imply adaptive evolutionary convergence. *RNA*, 3:792–806, 1997.

[34] M. Sipser. *Introduction to the theory of computation.* PWS Publishing Company, Boston, MA, 1997.

[35] G. Varani and W.H. McClain. The g x u wobble base pair. a fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep*, 1:18–23, 2000.

[36] W. Weaver and C. Shannon. *The mathematical theory of communication.* University of Illinois Press, Urbana, IL, 1949.

[37] M. Yarus and R. Knight. The scope of selection. In *The genetic code and the origin of life.* Landes Bioscience, Georgetown, TX, 2004.